Synopsis

Modern era of science is mainly driven by massive data. Today with numerous sensor networks, observatories, database and large scale simulations, the inflow of data per day is very large in terms of volume and diversity. Science has reached a point where discoveries are made from massive datasets rather than simple observations like during the ages of Kepler and Galileo. The broad range of observational data has poised many new challenges in the field of discovery science and data mining.

Uncertainties in observational data due to noise and missing values is one of the main challenges with scientific data analysis. For example, seismic data inherently has large scale noise in the data recorded from its sensors. It is often difficult to exactly identify the contributions from tectonic movement in seismic data to correct for those errors. Similarly data from climate sensors often have missing values, which makes it difficult to completely describe the physical phenomena underlying the process.

Noise in data may be introduced due to instrumental noise and environmental conditions. Besides the environment, the signal to noise ratio (SNR) of the instrument is one factor which decides the scale of noise in the data. In such cases it is often difficult to separate data from noise.

Data may have missing values mainly due to observational limitations. The limitations may arise because the source of observation is unavailable continuously or due to limitations in the design of the experiment. It is often difficult to decide whether the available data with missing values can still faithfully represent the underlying system.

To do good science, therefore it is essential to extract relevant properties with appropriate tools to study the system under observation. The general strategy in such situations is to transform the data into a different numerical form. This transformation will in principle be a unique representation of the original data. This method is called feature extraction. Feature extraction involves methods to transform raw data into numerical values called features. It transforms raw data into more informative signatures of the system. Feature extraction is critically important when it comes to clustering and classification of data and is the first transformation step for any kind of datamining.

The choice of feature extraction methods depends on the type of data used for a study. For time series data, the features extracted may include spectral properties, shape properties and even distance metrics with a standard template. Two dimensional data like images can be represented by similar spectral properties in a spatial domain. Each feature will represent the data in a different manner which will actually reflect its property in the respective domain. In this thesis, feature extraction methods, specifically for time series data which has inherent uncertainties is presented.

Time series data is one of the most common form of scientific data which are prone to noise and missing values. The feature extraction of noisy data is illustrated with EEG data (Electroencephalogram) and astronomical time series data is used for the feature extraction and analysis of data with missing values. Both these kinds of data have been growing to large volumes during the past decade and their analysis has wide spread applications.

We also develop a novel feature extraction method using wavelets and

its application to EEG data is discussed. The method can be applied to a wide variety of time series data which are similar in nature. EEG studies are currently of interest not only to neuroscientist but amongst physicist, mathematicians, computer scientists and even engineers. We use one of the least discussed property of wavelet asymmetry for the efficient feature extraction and detection of specific components from EEG signals with moderate amount of noise.

We also present the application of higher order spectral analysis mainly the Bispectral analysis as a feature extraction method for noisy data. Extracting information for signals embedded within noise levels of signals is a challenging task. Methods based on wavelets also have certain limitations when handling similar problems. Wavelet based thresholding mechanisms are useful to remove noise from signals, but is not useful to extract information embedded in highly noisy data. There may be non-linear interactions in signals which are feeble to identify with traditional linear methods. Bispectral analysis can be used to detect non-linear components in signals and can be used as features for classification problems. Using Bispectrum as a feature extraction method, we show its application to detect non-linear components in EEG signals

Feature extraction from partial data is another aspect we look into in this thesis. Observational limitations and data recording methods usually produce missing values in datasets. In astronomy, light curves are graphs that show the variation of an objects magnitude over time. It is particularly useful in the study of variable stars, periodic objects like RRLyrae and other similar celestial bodies. The study of light curves can yield a large amount

of information that underlies the physical processes of the objects. We look at different feature extraction methods and analyse their importance with sparse light curves from the Catalina Real-time Transient Survey and Kepler Space telescope. A class wise comparison of each feature and their ability for better classification is also studied.

The thesis is organized as follows:

Chapter 1 introduces the challenges in the data driven science and discusses about the general uncertainties in data. The need of feature extraction strategies and their general methods are also introduced.

Chapter 2 describes the novel method of feature extraction based on wavelets and its application to EEG signals. The application of Bispectrum to extract features from noisy signals is also presented.

Chapter 3 discusses the challenges with feature extraction from sparse light curves and analyses the relevance of existing feature extraction strategies. The methods to select important features for improved classification is also presented.

Chapter 4 summarizes the results and possible extensions to the study.

Publications

1. *A Wavelet based Algorithm for the Identification of Oscillatory Event-Related Potential Components* , Arun Kumar A., Ninan Sajeeth Philip, Vincent J. Samar, James A. Desjardins, Sidney J. Segalowitz, Accepted for publication in the Journal of Neuroscience Methods.

2. *Feature selection strategies for high dimensional datasets*, Ciro Donalek, Arun Kumar A., Ashish Mahabal et al., Proceedings of Scalable Machine Learning: Theory and Applications, IEEE BigData 2013

3. *Extraction of Information embedded in highly noisy data*, Arun Kumar A., Deepthi R., Ninan Sajeeth Philip.(In preparation)

4. *Characterization and Feature Selection of Kepler light curves*, Arun Kumar A., Ashish Mahabal, Matthew J Graham et. al. , (In preparation)

5. Feature space analysis of Kepler False Positives (In preparation), Arun Kumar A., Ashish Mahabal, S.G Gjorgosky et. al. , (In preparation)